DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 4 | Issue 1 Article 17

5-1-2005

On The Power Function Of Bayesian Tests With Application To Design Of Clinical Trials: The Fixed-Sample Case

Lyle Broemeling
University of Texas MD Anderson Cancer Center

Dongfeng Wu
Mississippi State University

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 163-171

1538 – 9472/05/$95.00

On The Power Function Of Bayesian Tests With Application To Design Of Clinical Trials: The Fixed-Sample Case

Lyle Broemeling
Department of Biostatistics and Applied Mathematics
University of Texas MD Anderson Cancer Center

Dongfeng Wu
Department of Mathematics and Statistics
Mississippi State University

Using a Bayesian approach to clinical trial design is becoming more common. For example, at the MD Anderson Cancer Center, Bayesian techniques are routinely employed in the design and analysis of Phase I and II trials. It is important that the operating characteristics of these procedures be determined as part of the process when establishing a stopping rule for a clinical trial. This study determines the power function for some common fixed-sample procedures in hypothesis testing, namely the one and two-sample tests involving the binomial and normal distributions. Also considered is a Bayesian test for multi-response (response and toxicity) in a Phase II trial, where the power function is determined.

Key words: Bayesian; power analysis; sample size; clinical trial

Introduction

The Bayesian approach to testing hypotheses is becoming more common. For example, in a recent review volume, see Crowley (2001), many contributions where Bayesian considerations play a prominent role in the design and analysis of clinical trials. Also, in an earlier Bayesian review (Berry & Stangl, 1996), methods are explained and demonstrated for a wide variety of studies in the health sciences, including the design and analysis of Phase I and II studies.

At our institution, the Bayesian approach is often used to design such studies. See Berry (1985,1987,1988), Berry and Fristed (1985), Berry and Stangl (1996), Thall and Russell (1998), Thall, Estey, and Sung (1999), Thall, Lee, and Tseng (1999), Thall and Chang (1999),and Thall et al. (1998), for some recent references where Bayesian ideas have been the primary consideration in designing Phase I and II studies. Of related interest in the design of a trial is the estimation of sample size based on Bayesian principles, where Smeeton and Adcock (1997) provided a review of formal decision-theoretic ideas in choosing the sample size.

Typically, the statistician along with the investigator will use information from previous related studies to formulate the null and alternative hypotheses and to determine what prior information is to be used for the Bayesian analysis. With this information, the Bayesian design parameters that determine the critical region of the test are given, the power function calculated, and lastly the sample size determined as part of the design process. In this study, only fixed-sample size procedures are used.

First, one-sample binomial and normal tests will be considered, then two-sample tests for binomial and normal populations, and lastly a test for multinomial parameters of a multi-response Phase II will be considered. For each test, the null and alternative hypotheses will be formulated and the power function determined. Each case will be illustrated with an example, where the power function is calculated for several values of the Bayesian design parameters.

Lyle Broemeling is Research Professor in the Department of Biostatistics and Applied Mathematics at the University of Texas MD Anderson Cancer Center. Email: lbroemel@mdanderson.org. Dongfeng Wu got her PhD from the University of California, Santa Barbara. She is an assistant professor at Mississippi State University.

Methodology

For the design of a typical Phase II trial, the investigator and statistician use prior information on previous related studies to develop a test of hypotheses. If the main endpoint is response to therapy, the test can be formulated as a sample from a binomial population, thus if Bayesian methods are to be employed, prior information for a Beta prior must be determined. However, if the response is continuous, the design can be based on a one-sample normal population. Information from previous related studies and from the investigator's experience will be used to determine the null and alternative hypotheses, as well as the other design parameters that determine the critical region of the test.

The critical region of a Bayesian test is given by the event that the posterior probability of the alternative hypothesis will exceed some threshold value. Once a threshold value is used, the power function of the test can be calculated. The power function of the test is determined by the sample size, the null and alternative hypotheses, and the above-mentioned threshold value.

Results

Binomial population

Consider a random sample from a Bernoulli population with parameters n and θ, where n is the number of patients and θ is the probability of a response. Let X be the number of responses among n patients, and suppose the null hypotheses is H: $\theta \le \theta_0$ versus the alternative A: $\theta > \theta_0$. From previous related studies and the experience of the investigators, the prior information for θ is determined to be Beta(a,b), thus the posterioir distribution of θ is Beta (x+a, n-x+b), where x is the observed number of responses among n patients. The null hypothesis is rejected in favor of the alternative when

$$\Pr[\theta > \theta_0 / x, n] > \gamma, \qquad (1)$$

where γ is usually some large value as .90, .95, or .99. The above equation determines the critical region of the test, thus the power function of the test is

$$g(\theta) = \Pr_{X/\theta} \{\Pr[\theta > \theta_0 / x, n] > \gamma\}, \qquad (2)$$

where the outer probability is with respect to the conditional distribution of X given θ. The power (2) at a given value of θ is interpreted as a simulation as follows:

(a) select (n, θ), and set S=0,

(b) generate a X~Binomial(n, θ),

(c) generate a θ~Beta(x+a, n-x+b),

(d) if Pr $[\theta > \theta_0 / x, n] > \gamma$, let the counter S =S+1, otherwise let S=S,

(e) repeat (b)-(d) M times, where M is 'large',

and

(f) select another θ and repeat (b)-(d).

The power of the test is thus S/M and can be used to determine a sample size by adjusting the threshold γ, the probability of a Type I error g(θ_0), and the desired power at a particular value of the alternative. The approach taken is fixing the Type I error at α and finding n so that the power is some predetermined value at some value of θ deemed to be important by the design team. This will involve adjusting the critical region by varying the value of the threshold γ. An example of this method is provided in the next section. The above hypotheses are one-sided, however it is easy to adjust the above testing procedure for a sharp null hypothesis.

Normal Population

Let N(θ, τ^{-1}) denote a normal population with mean θ and precision τ, where both are unknown and suppose we want to test the null hypothesis H: $\theta = \theta_0$ versus A: $\theta \neq \theta_0$, based on a random sample X of size n

with sample mean $\bar{X}$ and variance S^2. Using a non-informative prior distribution for θ and τ, the Bayesian test is to reject the null in favor of the alternative if the posterior probability P of the alternative hypothesis satisfies

$$P > \gamma, \text{ where} \quad (3)$$

$$P = D_2/D \quad (4)$$

and, $D = D_1 + D_2$.

It can be shown that

$$D_1 = \{\pi\Gamma(n/2)\}2^{n/2}\} / \{(2\pi)^{n/2} [\, n(\theta_0 - \bar{X})^2 + (n-1)\, S^2]^{n/2}\} \quad (5)$$

and

$$D_2 = \{(1-\pi)\Gamma((n-1)/2)\, 2^{(n-1)/2}\} / \{n^{1/2}(2\pi)^{(n-1)/2}[(n-1)\, S^2]^{(n-1)/2}\} \quad (6)$$

where π is the prior probability of the null hypothesis.

The power function of the test is

$$g(\theta,\tau) = \Pr_{X/\theta,\tau}[\, P > \gamma / \text{ n}, \bar{X}, S^2], \quad \theta \in R \text{ and } \tau > 0 \quad (7)$$

where P is given by (3) and the outer probability is with respect to the conditional distribution of X given θ and τ.

The above test is for a two-sided alternative, but the testing procedure is easily revised for one-sided hypotheses. This will be used to find the sample size in an example to be considered in a following section.

In the case when the null and alternative hypotheses are H: $\theta \le \theta_0$ and A: $\theta > \theta_0$ and the prior distribution for the parameters is $f(\theta,\tau) \propto 1/\tau$, where H is rejected in favor of A whenever

$$\Pr[\theta > \theta_0 / \text{n}, \bar{X}, S^2] > \gamma,$$

it can be shown that the power (size) of the test at θ_0 is 1-γ. Thus in this sense, the Bayesian and classical t-test are equivalent.

Two binomial populations

Comparing two binomial populations is a common problem in statistics and involves the null hypothesis H: $\theta_1 = \theta_2$ versus the alternative A: $\theta_1 \neq \theta_2$, where θ_1 and θ_2 are parameters from two Bernoulli populations. Assuming uniform priors for these two populations, it can be shown that the Bayesian test is to reject H in favor of A if the posterior probability P of the alternative hypothesis satisfies

$$P > \gamma, \text{ where} \quad (8)$$

$$P = D_2/D, \quad (9)$$

and $D = D_1 + D_2$. It can be shown that

$$D_1 = \{\pi \text{BC}(n_1 : x_1)\text{BC}(n_2 : x_2) \Gamma(x_1 + x_2 + 1)\Gamma(n_1 + n_2 - x_1 - x_2)\} \div \Gamma(n_1 + n_2 + 2),$$

where BC(n,x) is the binomial coefficient "x from n". Also, $D_2 = (1-\pi)(n_1+1)^{-1}(n_2+1)^{-1}$, where π is the prior probability of the null hypothesis. X_1 and X_2 are the number of responses from the two binomial populations with parameters (n_1, θ_1) and (n_2, θ_2) respectively. The alternative hypothesis is two-sided, however the testing procedure is easily revised for one-sided hypotheses.

In order to choose sample sizes n_1 and n_2, one must calculate the power function

$$g(\theta_1,\theta_2) = \Pr_{x_1,x_2/\theta_1,\theta_2}[P > \gamma / x_1, x_2, n_1, n_2], (\theta_1,\theta_2) \in (0,1)\,\text{x}\,(0,1) \quad (10)$$

where P is given by (9) and the outer probability is with respect to the conditional distribution of X_1 and X_2, given θ_1 and θ_2. As given above, (10) can be evaluated by a simulation procedure similar to that described in 3.1.

Two normal populations

Consider two normal populations with means θ_1 and θ_2 and precisions τ_1 and τ_2 respectively, and suppose the null and alternative hypotheses are H: $\theta_1 \leq \theta_2$ and A: $\theta_1 > \theta_2$ respectively. Assuming a non-informative prior for the parameters, namely $f(\theta_1,\theta_2,\tau_1,\tau_2) = 1/\tau_1\tau_2$, one can show that the posterior distribution of the two means is such that θ_1 and θ_2 are independent and θ_i /data ~ t(n_i-1, $\bar{X}_i$ n_i/S_i^2), where n_i is the sample size and $\bar{X}_i$ and S_i^2 are the sample mean and variance respectively.

That is, the posterior distribution of θ_i is a t distribution with n_i-1 degrees of freedom, mean $\bar{X}_i$, and precision n_i/S_i^2. It is known that $(\theta_i - \bar{X}_i)(n_i/S_i^2)^{1/2}$ has a Student's t-distribution with n_i-1 degrees of freedom.

Therefore the null hypothesis is rejected if

$$\Pr[\theta_1 > \theta_2/\text{data}] > \gamma . \quad (11)$$

Multinomial Populations

Consider a multinomial population with k categories and corresponding probabilities θ_i, i= 1,2,…,k, where $\sum_{i=1}^{i=k} \theta_i = 1$ and $0 < \theta_i < 1$ for i=1,2,…,k. Suppose there are n patients and that n_i belong to the i-th category.

The multinomial model is quite relevant to the Phase II trial where the k categories represent various responses to therapy. Let $\theta = (\theta_1,\theta_2,...,\theta_k)$, then if a uniform prior distribution is appropriate, the posterior distribution is

$$f(\theta/\text{ data}) \propto \prod_{i=1}^{i=k} \theta_i^{n_i}, \quad \sum_{i=1}^{i=k} \theta_i = 1, \text{ and } 0 < \theta_i < 1 \text{ for i=1,2,…,k.} \quad (12)$$

and the distribution is Dirichlet $(n_1+1, n_2+1,..., n_k+1)$.

A typical hypothesis testing problem, see [14], is given by the null hypothesis (k=4), where

$$\text{H: } \theta_1 + \theta_2 \leq k_{12} \text{ or } \theta_1 + \theta_3 \geq k_{13}$$

versus the alternative

$$\text{A: } \theta_1 + \theta_2 > k_{12} \text{ and } \theta_1 + \theta_3 < k_{13} .$$

The null hypothesis states that the response rate $\theta_1 + \theta_2$ is less than some historical value or that the toxicity rate $\theta_1 + \theta_3$ is greater than some historical value k_{13}. The null hypothesis is rejected if the response rate is larger than the historical or the toxicity rate is too low compared to the historical.

$$\Pr[A/\text{data}] > \gamma \quad (13)$$

where γ is some threshold value. This determines the critical region of the test, thus the power function is

$$g(\theta) = \Pr_{n/\theta}\{ \Pr[A/\text{ data}] > \gamma \}, \quad (14)$$

where the outer probability is with respect to the conditional distribution of

$$n = (n_1, n_2,..., n_k) \text{ given } \theta .$$

The power function will be illustrated for the multinomial test of hypothesis with a Phase I trial, where response to therapy and toxicity are considered in designing the trial.

Examples

The above problems in hypothesis testing are illustrated by computing the power function of some Bayesian tests that might be used in the design of a Phase II trial.

One-Sample Binomial

No prior information

Consider a typical Phase II trial, where the historical rate for toxicity was determined as .20. The trial is to be stopped if this rate exceeds the historical value. See Berry (1993) for a good account of Bayesian stopping rules in clinical trials. Toxicity rates are carefully defined in the study protocol and are based on the NCI list of toxicities. The null and alternative hypotheses are given as

$$\text{H: } \theta \le .20 \text{ and A: } \theta > .20, \qquad (15)$$

where θ is the probability of toxicity. The null hypothesis is rejected if the posterior probability of the alternative hypothesis is greater than the threshold value γ.

The power curve for the following scenarios will be computed (see Equation 2), with sample sizes n = 125, 205, and 500, threshold values γ = .90, .95, .99, M=1000, and null value θ_0 = .20.

It is seen that the power of the test at $\theta = .30$ and γ = .95, is .841, .958, and .999 for n = 125, 205, and 500, respectively.

Note that for a given N and γ, the power increases with θ and for given N and θ, the power decreases with γ, and for given γ and θ, the power of course increases with N.

The Bayesian test behaves in a reasonable way. For the conventional type I error of .05, a sample size of N=125 would be sufficient to detect the difference .3 versus .2 with a power of .841. It is interesting to note that the usual binomial test, with alpha = .05 and power .841, requires a sample of size 129 for the same alternative value of θ. For the same alpha and power, one would expect the Bayesian (with a uniform prior for θ) and the binomial tests to behave in the same way in regard to sample size.

With prior information

Suppose the same problem is considered as above, but prior information is available with 50 patients, 10 of whom have experienced toxicity. The null and alternative hypotheses are as above, however the null is rejected whenever

$$\Pr[\theta > \phi / \text{x, n}] > \gamma, \qquad (16)$$

where θ is independent of $\phi \sim$ Beta(10,40). This can be considered as a one-sample problem where a future study is to be compared to a historical control.

As above, compute the power function (see Table 2) of this Bayesian test with the same sample sizes and threshold values in Table 1. The power of the test is .758, .865, and .982 for θ = .4 for N= 125, 205, and 500, respectively. This illustrates how important is prior information in testing hypotheses. If the hypothesis is rejected with the critical region

$$\Pr[\theta > .2 / \text{x, n}] > \gamma, \qquad (17)$$

the power (Table 1) will be larger than the corresponding power (Table 2) determined by the critical region (16), because of the additional posterior variability introduced by the historical information contained in ϕ. Thus, larger sample sizes are required with (16) to achieve the same power as with the test given by (17).

Table 1. Power function for H versus A, N=125,205,500.

θ \ γ	.90	.95	.99
0	0,0,0	0,0,0	0,0,0
.1	0,0,0	0,0,0	0,0,0
.2	.107,.099,.08	.047,.051,.05	.013,.013,.008
.3	.897,.97,1	.841,.958,.999	.615,.82,.996
.4	1,1,1	1,1,1	.996,1,1
.5	1,1,1	1,1,1	1,1,1
.6	1,1,1	1,1,1	1,1,1
.7	1,1,1	1,1,1	1,1,1
.8	1,1,1	1,1,1	1,1,1
.9	1,1,1	1,1,1	1,1,1
1.0	1,1,1	1,1,1	1,1,1

Table 2. Power function for H versus A, N=125,205,500.

θ \ γ	.90	.95	.99
0	0,0,0	0,0,0	0,0,0
.1	0,0,0	0,0,0	0,0,0
.2	.016,.001,.000	.002,.000,.000	.000,.000,.000
.3	.629,.712,.850	.362,.374,.437	.004,.026,.011
.4	.996,.999,1	.973,.998,1	.758,.865,.982
.5	1,1,1	1,1,1	.999,1,1
.6	1,1,1	1,1,1	1,1,1
.7	1,1,1	1,1,1	1,1,1
.8	1,1,1	1,1,1	1,1,1
.9	1,1,1	1,1,1	1,1,1
1.0	1,1,1	1,1,1	1,1,1

Two Binomial Populations

The case of two binomial populations was introduced in section 4.2, where equation (10) gives the power function for testing H: $\theta_1 = \theta_2$ versus the alternative A: $\theta_1 \neq \theta_2$.

In this example, let $n_1 = 20 = n_2$ be the sample sizes of the two groups and suppose the prior probability of the null hypotheses is π = .5. The power at each point (θ_1, θ_2) is calculated via simulation, using equation (10) with γ = .90. Table 3 lists the power function for this test.

When the power is calculated with the usual two-sample, two-tailed, binomial test with alpha = .013, sample sizes $n_1 = 20 = n_2$, and (θ_1, θ_2) = (.3, .9), the power is .922, which is almost equivalent to the above Bayesian test. This is to be expected, because we are using a uniform prior density for both Bernoulli parameters. It is not too uncommon to have two binomial populations in a Phase II setting, where θ_1 and θ_2 are response rates to therapy.

Table 3. Power for Bayesian Binomial Test.

θ_1 \ θ_2	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
.1	.004	.032	.135	.360	.621	.842	.958	.992	1	1
.2	.031	.011	.028	.106	.281	.536	.744	.913	.997	1
.3	.171	.028	.006	.029	.107	.252	.487	.767	.961	1
.4	.368	.098	.025	.013	.028	.075	.244	.542	.847	.999
.5	.619	.289	.100	.022	.007	.017	.108	.291	.640	.981
.6	.827	.527	.237	.086	.035	.005	.027	.116	.357	.882
.7	.950	.775	.464	.254	.113	.037	.013	.049	.171	.587
.8	.996	.928	.768	.491	.316	.132	.028	.010	.040	.205
.9	1	.996	.946	.840	.647	.359	.156	.037	.006	.014
1	1	1	1	1	.984	.873	.567	.200	.017	.000

A Phase II trial with toxicity and response rates

With Phase II trials, response to therapy is usually taken to be the main endpoint, however in reality one is also interested in the toxicity rate, thus it is reasonable to consider both when designing the study. Most Phase II trials are conducted not only to estimate the response rate, but to learn more about the toxicity. In such a situation, the patients can be classified by both endpoints as follows:

Table 4. Number of and Probability of Patients by Response and Toxicity.

Response \ Toxicity	Yes	No
Yes	(n_1, θ_1)	(n_2, θ_2)
No	(n_3, θ_3)	(n_4, θ_4)

Let the response rate be $\theta_r = \theta_1 + \theta_2$ and the rate of toxicity be $\theta_t = \theta_1 + \theta_3$, where θ_1 is the probability a patient will experience toxicity and respond to therapy, and n_1 is the number of patients who fall into that category. Following Petroni and Conoway (2001), let the null hypothesis be

$$\text{H: } \theta_r \le \theta_{r0} \text{ or } \theta_t \ge \theta_{t0}$$

and the alternative hypothesis be

$$\text{A: } \theta_r > \theta_{r0} \text{ and } \theta_t < \theta_{t0},$$

where θ_{r0} and θ_{t0} are given and estimated by the historical rates in previous trials.

Table 5. Power of Bayesian Multinomial Test.

	θ_t	.2	.3	.4	.5
θ_r					
.2		.000	.000	.000	.000
.3		.000	.000	.000	.000
.4		.070	.002	.000	.000
.5		.600	.114	.000	.000
.6		.794	.154	.000	.000
.7		.818	.158	.000	.000
.8		.822	.084	.000	.000

In this example, let θ_{r0} = .40 and θ_{t0} = .30. That is, the alternative hypothesis is that the response rate exceeds .40 and the toxicity rate is less than .30, and the null is rejected in favor of the alternative if the latter has a posterior probability in excess of γ. Table 5 gives the power for n=100 patients and threshold γ = .90.

From above, the power of the test is .818 when (θ_r, θ_t) = (.7, .2), and the test behaves in a reasonable way. When the parameter values are such that the response rate is in excess of .40 and the toxicity rate is less than or equal to .30, the power is higher, relative to those parameter values when the null hypothesis is true.

Conclusion

We have provided a way to assess the sampling properties of some Bayesian tests of hypotheses used in the design and analysis of Phase II clinical trials.

The one-sample binomial scenario is the most common in a Phase II trial, where the response to therapy is typically binary. We think it is important to know the power function of a critical region that is determined by Bayesian considerations, just as it is with any other test.

The Bayesian approach has one major advantage and that is prior information, and when this is used in the design of the trial, the power of the test will be larger then if prior information had not been used.

We have confined this investigation to the fixed-sample case, but will seek to expand the results to the more realistic situation where Bayesian sequential stopping rules will be used to design Phase II studies.

References

Berry D. A. (1985). Interim analysis in clinical trials: Classical versus Bayesian approaches. *Statistics in Medicine*, *4*, 521-526.

Berry D. A. (1987). Interim analysis in clinical trials: the role of the likelihood principal. *The American Statistician*, *41*, 117-122.

Berry D. A. (1988). Interim analysis in clinical research. *Cancer Investigations*, *5*, 469-477.

Berry D. A. (1993). A case for Bayesianism in clinical trials (with discussion). *Statistics in Medicine*, *12*, 1377-1404.

Berry D. A., & Fristed, B. (1985). *Bandit problems. Sequential allocation of experiments*. New York: Chapman-Hall.

Berry D. A., & Stangl D. K., (1996). Bayesian methods in health-related research,. *Bayesian Biostatistics*. (In D. A. Berry, & D. Stangl, eds.) New York: Marcel Dekker Inc., p. 3 – 66.

Crowley J. (2001). *Handbook of statistics in clinical oncology*. New York: Marcel Dekker Inc.

Petroni G. R., & Conoway M. R. (2001). Designs based on toxicity and response. (In J. Crowley, ed.) *Handbook of Statistics in Clinical Oncology*. New York: Marcel-Dekker Inc., p. 105 – 118.

Smeeton N. C., & Adcock C. J. (1997, special issue). (eds.) Sample size determination. *The Statistician*, *4*.

Thall P. F, Simon R., Ellenberg S. S., & Shrager R. (1988). Optimal two-stage designs for clinical trials with binary responses. *Statistics in Medicine*, *71*, 571-579.

Thall P. F., & Russell K. E. (1998). A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics*, *54*, 251-264.

Thall P. F., Estey E. H., & Sung H. G. (1999). A new statistical method for dose-finding based on efficacy and toxicity in early phase clinical trials. *Investigational New Drugs*, *17*, 155-167.

Thall P. F., Lee J. J., & Tseng C. H. (1999). Accrual strategies for Phase I trials with delayed patient outcome. *Statistics in Medicine*, *18*, 1155-1169.

Thall P. F., & Cheng S. C. (1999) Treatment comparisons based on two-dimensional safety and efficacy alternatives in oncology trials. *Biometrics*, *55*, 746-753.